# Introduction to Dataset

Interesting Dataset?

In order to compare the performance of different machine learning algorithms, two datasets, UCI’s Breast Cancer Wisconsin Data Set and Caltech 101 Data Set, have been deliberately chosen. Even though both datasets deal with image classification, one of these datasets (UCI dataset) has been translated into categorical and numerical data, while the other stays in its raw pixel form. In addition, only the important features are included in the UCI dataset dataset and has a binary classification output. This contrast with the Caltech dataset which has 101 categories output. In other words, this paper tries to compare the capacity of various supervised learning algorithm on recognizing noises on a “easy” (the feature extracted and less noisy binary UCI dataset) versus “hard” (the raw and multi-categorical Caltech dataset) problem.

The Caltech 101 dataset has approximately 10k instances, while the UCI Breast Cancer Dataset contains 569 entries. In order to make the comparison, the Caltech 101 dataset has been reduced to 20 categories with approximately 2k data points. This is also done since the time to load and train algorithms have been too long to neglect, taking as much as 4 hours to run.

Considering the number of categories for the Caltech dataset (around 100 images per category), this paper will assume it is almost fair to make the comparison. (It is not easy to find another dataset with the right number of instances.) On other hand, as the result unfold, if there is a disparaging difference between the performance, there would be a discussion on whether the number of training data could have affected the performance. Nonetheless, the difference in size would also enable a comparison between training and testing time for various algorithms.

A Note on Preprocessing and Grid Search (w/ Cross Validation)

To facilitate fair and comparable results, the same preprocessing steps are performed to the datasets. In particular, the images in the Caltech 101 dataset are resized, converted to grayscale (to make training faster), and feature extracted with histogram of oriented gradients. The hyperparameters are chosen by referencing other online repositories that use the Caltech 101 dataset. In addition, the labels for the Caltech dataset are one hot encoded. On the other hand, the UCI dataset is normalized (excluding the classification result.)

For each algorithm, some form of grid search to find the best hyperparameters is used to ensure that result is not biased from the lack of hyperparameter tuning. In the case of most algorithms, the two most “important” hyperparameters are chosen and plotted to visualize performance. The arbitrary choice of these “important” hyperparameters will be explained for each algorithm. Also, two are chosen so it is easy to graph (performance versus grid parameter 1, and one plotted line for each grid parameter 2) and trains in a shorter amount of time as grid search runs an exhaustive search.

# k-Nearest Neighbor

UCI Breast Cancer Dataset

The Scikit-Learn framework supports some of the most common distance metrics, including metrics optimized for two-dimensional vector space, integer-value vector space, boolean-value vector space (where input data are in the form of boolean values, not the output), and real-value vector space which best fit the UCI dataset.

